Assignments

Contents

This page will contain all the assignments you submit for the class.

Instructions for all assignments

I want you to submit your assignment as a PDF, so I can keep a record of what the code looked like that day. I also want you to include your answers on your personal GitHub website. This will be good practice for editing your website and it will help you produce something you can keep after the class is over.

1. Download the Assignment1.Rmd file from Canvas. You can use this as a template for writing your answers. It's the same as what you can see on my website in the Assignments tab. Once we're done with this I'll edit the text on the website to include the solutions.
2. On RStudio, open a new R script in RStudio (File > New File > R Script). This is where you can test out your R code. You'll write your R commands and draw plots here.
3. Once you have finalized your code, copy and paste your results into this template (Assignment 1.Rmd). For example, if you produced a plot as the solution to one of the problems, you can copy and paste the R code in R markdown by using the ``` ```{r} ``` ``` command. Answer the questions in full sentences and Save.
4. Produce a PDF file with your answers. To do this, knit to PDF (use Knit button at the top of RStudio), locate the PDF file in your docs folder (it's in the same folder as the Rproj), and submit that on on Canvas in Assignment 1.
5. Build Website, go to GitHub desktop, commit and push. Now your solutions should be on your website as well.

Assignment 1

Collaborators: Theo Athanitis.

Problem 1

Install the datasets package on the console below using `install.packages("datasets")`. Now load the library.

```
#install.packages("datasets")
library(datasets)

#install.packages("knitr")
library(knitr) #used for knitting to a pdf
```

Load the USArrests dataset and rename it `dat`. Note that this dataset comes with R, in the package datasets, so there's no need to load data from your computer. Why is it useful to rename the dataset?

```
dat <- USArrests
```

It is useful to rename the dataset as it creates a copy of the dataset stored in this newly created variable that can now be modified without changing the original, such as adding a new column for states. Additionally, renaming it can make it easier to call/ reference in later functions.

Problem 2

Use this command to make the state names into a new variable called State.

```
dat$state <- tolower(rownames(USArrests))
```

This dataset has the state names as row names, so we just want to make them into a new variable. We also make them all lower case, because that will help us draw a map later - the map function requires the states to be lower case.

List the variables contained in the dataset `USArrests`.

```
names(dat)
```

```
## [1] "Murder"   "Assault"  "UrbanPop" "Rape"     "state"
```

The variables contained in this dataset are Murder, Assault, UrbanPop, Rape, and state.

Problem 3

What type of variable (from the DVB chapter) is `Murder`?

Answer: Quantitative- the values of Murder are numerical values with measurment units as they record the amount of Murder arrests per 100,000 people in each state.

What R Type of variable is it?

Answer: The variable Murder itself is of the type character (as shown using typeof('Murder')), but the Murder values for each state are numeric doubles (as shown with dat[1,1])

```
typeof('Murder')
```

```
## [1] "character"
```

```
typeof(dat[1,1])
```

```
## [1] "double"
```

Problem 4

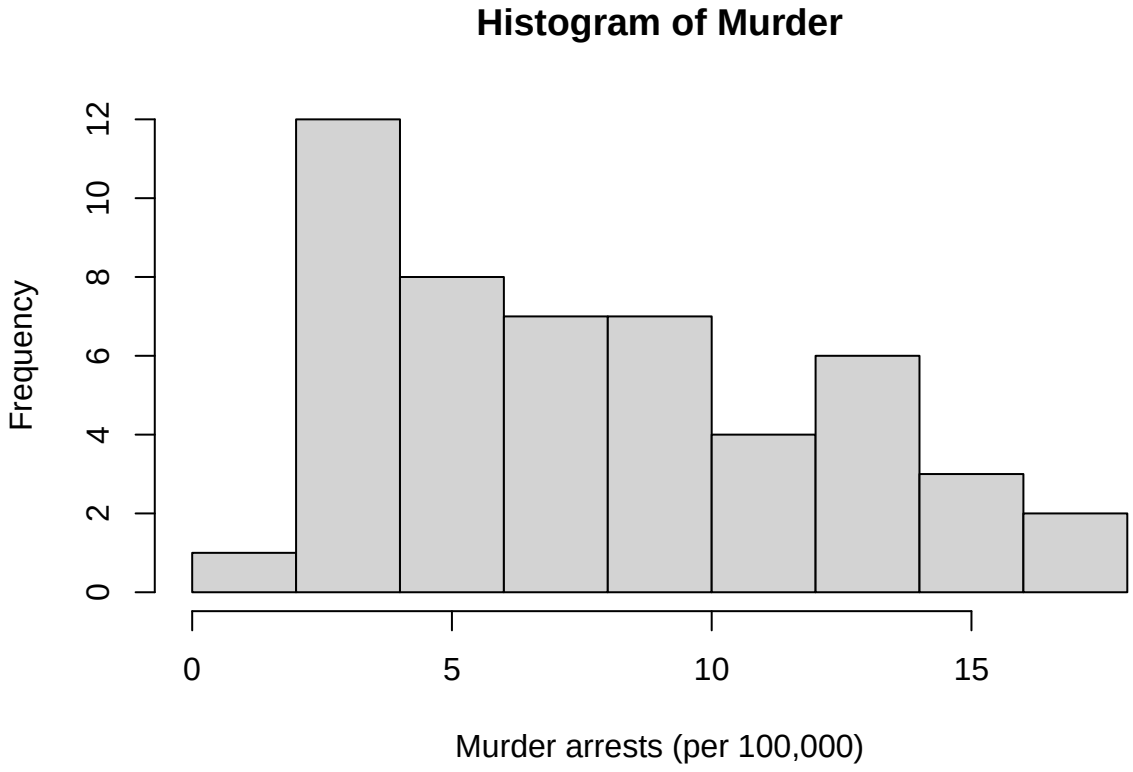
What information is contained in this dataset, in general? What do the numbers mean?

Answer: This dataset contains data on violent crime rates and urban area populations for each of the 50 states in 1973. The variables and their values contained are 1. the number of murder arrests per 100,000 people, 2. the number of assault arrests per 100,000 people, 3. the percentage of the population that lives in an urban environment, and 4. the number of rape arrests per 100,000 people for every US state.

Problem 5

Draw a histogram of `Murder` with proper labels and title.

```
hist(dat$Murder, main="Histogram of Murder", xlab="Murder arrests (per 100,000)", ylab="Frequency")
```



Problem 6

Please summarize `Murder` quantitatively. What are its mean and median? What is the difference between mean and median? What is a quartile, and why do you think R gives you the 1st Qu. and 3rd Qu.?

```
mean(dat$Murder)
```

```
## [1] 7.788
```

```
median(dat$Murder)
```

```
## [1] 7.25
```

```
quantile(dat$Murder)
```

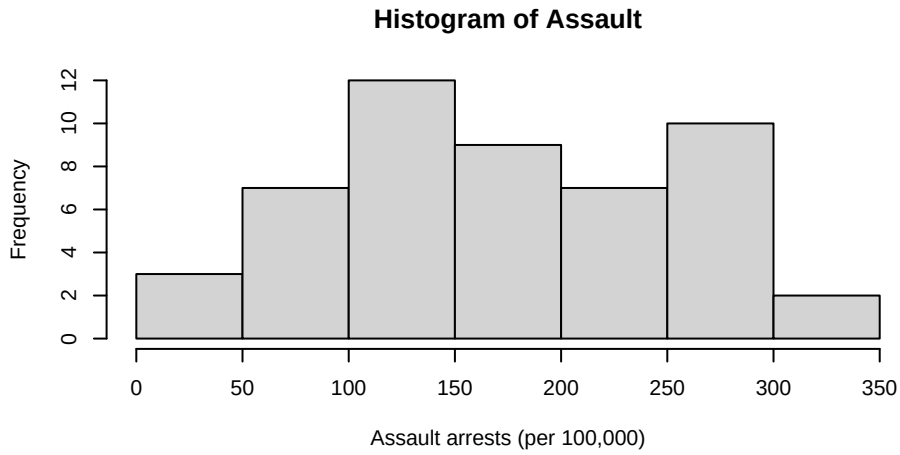
```
##     0%    25%    50%    75%   100%
##  0.800  4.075  7.250 11.250 17.400
```

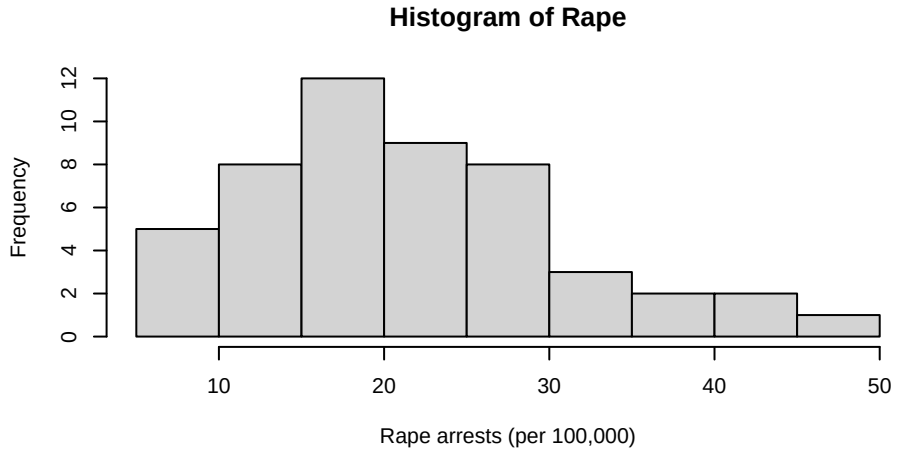
The mean of Murder is 7.788 and the median of Murder is 7.25. Mean is the sum of all of the states' murder arrests per 100,000 people divided by 50 (the number of states) also known as the average, while median is the middle value, where half of the value are greater than and half are less than the median, of the states' murder arrests per 100,000 people values. A quartile is one of three values which divide the dataset into four equal divisions. R likely only provides the 1st and 3rd quartiles (although I did not find a quartile function that operates this way and instead utilized the quantile function) as the 2nd quartile is the same as the median, therefore making the 1st and 3rd much more useful as they are more likely still unknown in comparison to the 2nd quartile when utilizing this function.

Problem 7

Repeat the same steps you followed for `Murder`, for the variables `Assault` and `Rape`. Now plot all three histograms together. You can do this by using the command `par(mfrow=c(3,1))` and then plotting each of the three.

```
par(mfrow=c(3,1))
hist(dat$Murder, main="Histogram of Murder", xlab="Murder arrests (per 100,000)", ylab="Frequency")
hist(dat$Assault, main="Histogram of Assault", xlab="Assault arrests (per 100,000)", ylab="Frequency")
hist(dat$Rape, main="Histogram of Rape", xlab="Rape arrests (per 100,000)", ylab="Frequency")
```





What does the command par do, in your own words (you can look this up by asking R `?par`)?

Answer: The par command is used to modify the manner in which graphs are displayed by finding, modifying, or setting the parameters of graphs. One functionality of par includes the ability to show multiple graphs together in the same graphic as shown here. In this instance, the mfrow=c(3,1) parameter is a vector with subplots of 1 in length (row) and 3 in depth (column), to create the stacked graphs the function above produces.

What can you learn from plotting the histograms together?

Answer: By plotting the histograms together, it is easier to see the distributional differences between the different variables. By organizing the histograms this way, it can be seen that the frequency by state for murder and rape arrests per 100,000 people are skewed to the left, while the frequency of assult arrests by state are more evenly distributed. While these histograms are not directly comparable because of their different scales, skewdness can still be recognized across them and this display makes it easier to recognize.

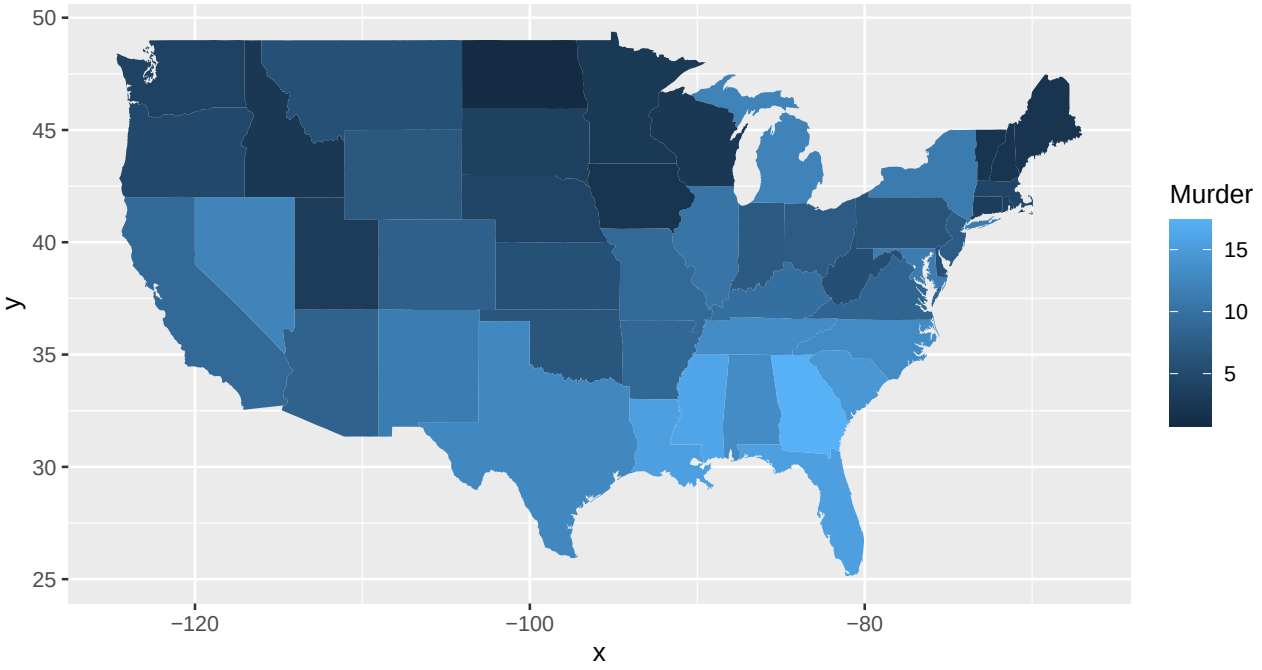
Problem 8

In the console below (not in text), type `install.packages("maps")` and press Enter, and then type `install.packages("ggplot2")` and press Enter. This will install the packages so you can load the libraries.

Run this code:

```
library('maps')
library('ggplot2')


ggplot(dat, aes(map_id=state, fill=Murder)) +
  geom_map(map=map_data("state")) +
  expand_limits(x=map_data("state")$long, y=map_data("state")$lat)
```



What does this code do? Explain what each line is doing.

Answer: First, the libraries for maps and ggplots are loaded. Next the ggplot function is called with first the parameter of dat as the dataset for the plot. The next set of parameters is for the aesthetic mapping for the plot basing the map_id for the values on the state variables and the fill/ color of that fill based on the Murder variable value for that representative map_id, as just defined. The fifth line further modifies the aesthetic mapping by defining the map for establishing the coordinate locations to display/divide the

states and their fills. Lastly, the sixth line further modifies the aesthetic mapping by defining the x and y limits of this graphic based on the x and y limits from the state positional variables using their latitude and longitudinal values to ensure that all values are displayed.

These last three lines together are creating a map of the Murder arrests per 100,000 people for each state by scaling the color of each state on a map of the United States to represent the degree of this amount in comparion to the other US states.

Assignment 2

Problem 1: Load data

Set your working directory to the folder where you downloaded the data.

```
#setwd("/Users/hwasser/Documents/Penn/4 Fourth Year/CRIM 250/Assignment 2")
```

Read the data

```
dat <- read.csv(file = 'dat.nsduh.small.1.csv')
```

What are the dimensions of the dataset?

```
dim(dat)
```

```
## [1] 171   7
```

There are 171 rows (excluding the row containing the column names) and 7 columns

Problem 2: Variables

Describe the variables in the dataset.

```
names(dat)
```

```
## [1] "mjage"     "cigage"    "iralcage"  "age2"      "sexatract" "speakengl"
## [7] "irsex"
```

• mjage is the age that respondents reported first trying marijuana or hashish • cigage is the age that respondents reported first started smoking cigarettes everyday • iralcage is the age that respondents reported first trying alcohol • age2 is a variable that represents an age bracket corresponding to the consistency checked age from respondents reported age from the beginning of the survey, end of the survey, and their birth date • sexatract is a variable representing categories of sexual attraction ranging from only attraction to the opposite sex to only attraction to the same sex reported by the respondent • speakengl is a variable that represents how well a respondent reports their ability to speak English with categories of very well, well, not well, and not at all • irsex is a variable that represents the respondents reported gender with 1 as male and 2 as female

What is this dataset about? Who collected the data, what kind of sample is it, and what was the purpose of generating the data?

This dataset is a sample from the 2019 National Survey for Drug Use and Health that focuses on the data collected on tobacco, alcohol, and drug use in the United States, along with the demographic information for these respondents, and is a stratified random sample. This data is collected by the Substance Abuse and Mental Health Services Administration (SAMHSA), which is an agency in the U.S. Department of Health and Human Services (HHS), in order to provide support for prevention/ treatment programs, monitor substance use trends in the United States, and ultimately inform public health policy by estimating the need for treatment in the US based on this data.

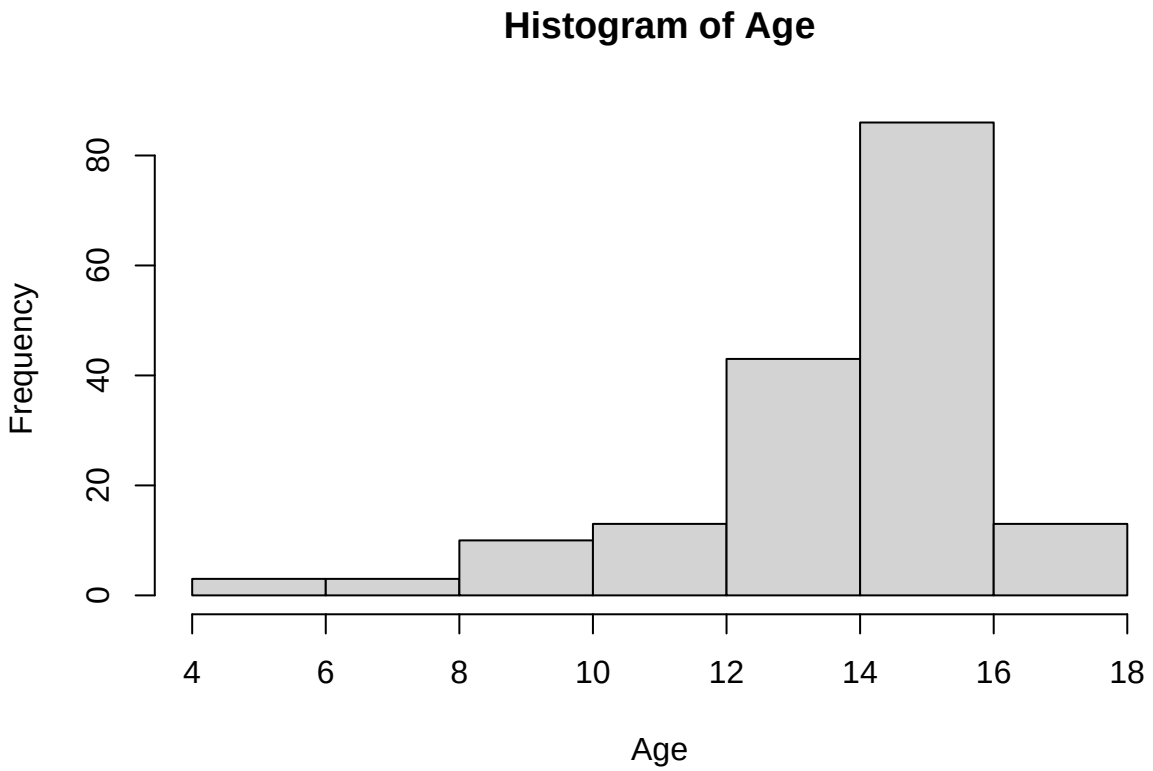
Problem 3: Age and gender

What is the age distribution of the sample like? Make sure you read the codebook to know what the variable values mean.

```
table(dat$age2)
```

```
##
##  4  6  7  8  9 10 11 12 13 14 15 16 17
##  2  1  1  2  7  3  6  7 27 16 62 24 13
```

```
hist(dat$age2, main = "Histogram of Age", xlab = "Age")
```



This histogram shows a clear skew to the right with distribution with the largest number of respondents being between 39-45 years old, with very few respondents under the age of 20 years old.

Do you think this age distribution representative of the US population? Why or why not?

I do not believe that this age distribution is representative of the US population as there were no respondents (0%) being under 15 years of age and only 7.6% of respondents being 65 or older. The actual US population is 18.37% for ages 0-14 and 16.63% for 65 and older, which is clearly not represented by the respondents in this survey, however including the younger demographic would not be useful in this data collection as they would likely be unable to provide answers to the majority of these questions about drug and alcohol use. Furthermore, this sample has large peaks, as shown my the high frequency for the 39-45 year old respondents, however the US population distribution is much more consistent in this middle section with no peaks.

Is the sample balanced in terms of gender? If not, are there more females or males?

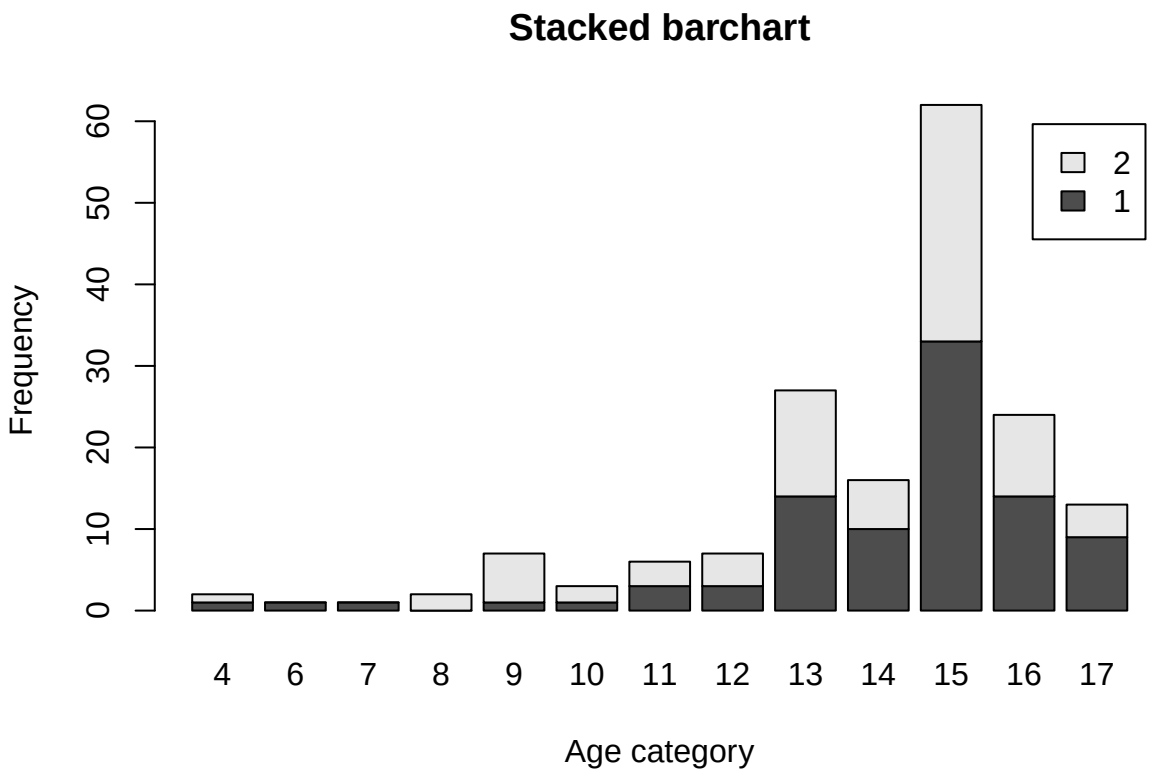
```
table(dat$irsex)
```

```
##
##  1  2
## 91 80
```

There are 91 respondents for 1 and 80 respondents for 2, so there are more males than females in the sample.

Use this code to draw a stacked bar plot to view the relationship between sex and age. What can you conclude from this plot?

```
tab.agesex <- table(dat$irsex, dat$age2)
barplot(tab.agesex,
        main = "Stacked barchart",
        xlab = "Age category", ylab = "Frequency",
        legend.text = rownames(tab.agesex),
        beside = FALSE) # Stacked bars (default)
```

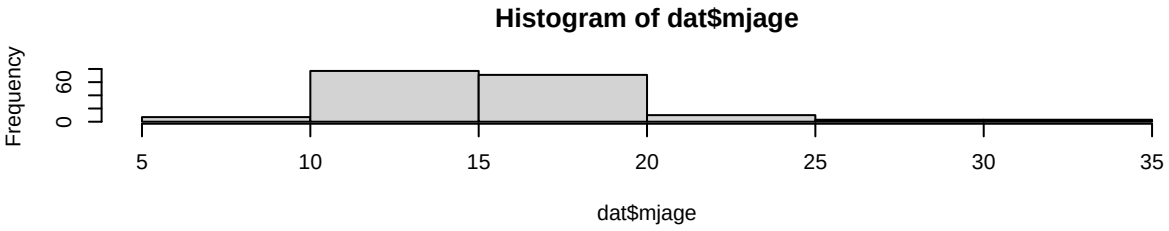


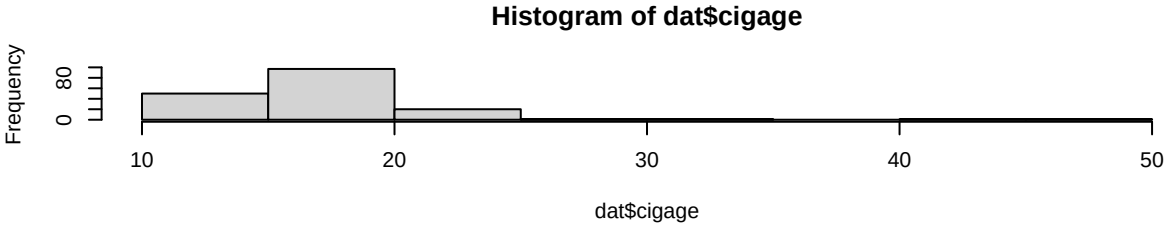
From this stacked bar graph, it appears that there seems to be a fairly even split between male and female respondents on average for the age categories. Looking at it more carefully though, it can be seen that the bins 6, 7, 14, 16, and 17 have at least slightly more male respondents than female respondents, while bins 8 and 9 have more female than male respondents.

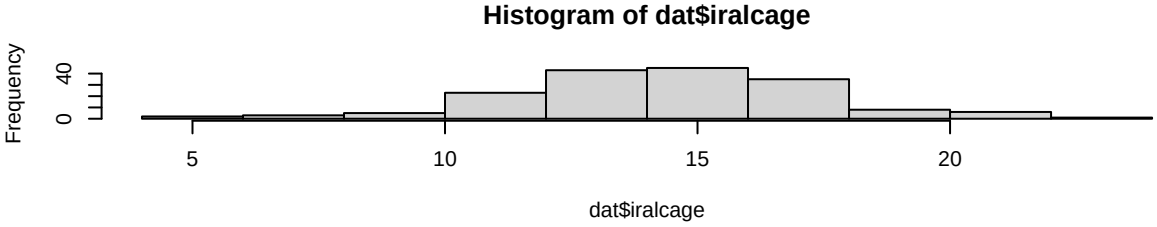
Problem 4: Substance use

For which of the three substances included in the dataset (marijuana, alcohol, and cigarettes) do individuals tend to use the substance earlier?

```
par(mfrow=c(3,1))
hist(dat$mjage)
hist(dat$cigage)
hist(dat$iralcage)
```








```
table(dat$mjage)
```

```
##
##  7  9 10 11 12 13 14 15 16 17 18 19 20 21 22 25 27 30 32 33 35
##  1  4  2  7 10 16 22 22 28 16 16  4  7  6  2  2  2  1  1  1  1
```

```
table(dat$cigage)
```

```
##
## 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 27 35 45 50
##  1  1  3 10 10 25 25 20 31 11 10  6  5  4  1  4  1  1  1  1
```

```
table(dat$iralcage)
```

```
##
##  5  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 23
##  2  1  2  1  4  4 19 21 22 19 26 12 23  6  2  6  1
```

It appears that of the three substances, individuals tend to try alcohol earliest. While the early age of use is similar across these substances, alcohol has the youngest responses with 2 at age 5, in comparison to 7 and 10 for cigarettes and marijuana, as well as having similar numbers of respondents for the following years as marijuana, demonstrating these are not pure outliers, making it the substance that individuals tend to use earlier. For each substance, the highest reported age frequency was 16, demonstrating the early similarity between them, but alcohol has more respondents on the younger end of this spectrum than marijuana or cigarettes, as well as not having any respondents greater than 23, in comparison to 35 and 50 for the other substances.

Problem 5: Sexual attraction

What does the distribution of sexual attraction look like? Is this what you expected?

```
table(dat$sexatract)
```

```
##
##   1   2   3   4   5   6  99
## 136  16   9   3   3   1   3
```

The distribution of sexual attraction is strongly skewed to the left towards only attraction to the opposite sex. Honestly this is not what I expected, as I would say that at least a 30-40% of the people I know in my generation identify as not straight, but I should have recognized the generational divide and as the survey respondents ages are distributed more on the older side, this should have been more expected. When comparing these numbers to national surveys on sexual attraction, especially ones with larger amounts of older

What is the distribution of sexual attraction by gender?

```
table(dat$sexatract, dat$irsex)
```

```
##
##       1  2
##   1  82 54
##   2   3 13
##   3   0  9
##   4   1  2
##   5   2  1
##   6   1  0
##   99  2  1
```

When comparing sexual attraction by gender, it appears that there are more women who identify as mostly attracted to the same sex and equally attracted to males and females than there are male respondents for either of these categories. Compensating for these differences, there are more males that responded that they are only attracted to the opposite sex than females.

Problem 6: English speaking

What does the distribution of English speaking look like in the sample? Is this what you might expect for a random sample of the US population?

```
table(dat$speakengl)
```

```
##
##   1   2   3
## 161   8   2
```

The vast majority of respondents report speaking English very well with 161 respondents, and then only 8 respondents for well and 2 for not well. This is consistent with what would be expected from a random sample of the US population. The US census reports that only 8.6% of the population "does not have a firm grasp of the English language", making the 94% of respondents in the survey stating very well is consistent with this data.

Are there more English speaker females or males?

```
table(dat$speakengl, dat$irsex)
```

```
##
##      1  2
##   1 84 77
##   2  7  1
##   3  0  2
```

There are more English speakers that are male with 91 responding either very well or well, with only 78 females responding to these categories, however this can be accounted for by the fact that more males responded to the survey in general. 100% of male respondents reported that they spoke English very well or well, and 97.5% of female male respondents reported that they spoke English very well or well.

Midterm

```
#setwd("/Users/hwasser/Documents/Penn/4 Fourth Year/CRIM 250/Exam 1")
dat <- read.csv(file = 'fatal-police-shootings-data.csv')
```

Problem 1 (10 points)

a. Describe the dataset. This is the source: https://github.com/washingtonpost/data-police-shootings . Write two sentences (max.) about this.

This dataset contains recordings of all fatal shootings by police in the United States since January 1st 2015 to the present (last updated 4 days ago) compiled by the Washington Post from news reports, law eenforcment sites, social media, and independent databases. For each fatal shooting, information about the victim and the manner in which they were treated by police is included, such as demographic information, whether or not the officer was wearing a body camera, the location, and the 'threat factor'.

b. How many observations are there in the data frame?

```
dim(dat)
```

```
## [1] 6594   17
```

There are 6594 observations of 17 variables.

c. Look at the names of the variables in the data frame. Describe what "body_camera", "flee", and "armed" represent, according to the codebook. Again, only write one sentence (max) per variable.

```
names(dat)
```

```
##  [1] "id"                     "name"
##  [3] "date"                   "manner_of_death"
##  [5] "armed"                  "age"
##  [7] "gender"                 "race"
##  [9] "city"                   "state"
## [11] "signs_of_mental_illness" "threat_level"
## [13] "flee"                   "body_camera"
## [15] "longitude"              "latitude"
## [17] "is_geocoding_exact"
```

```
#table(dat$body_camera)
#table(dat$flee)
#table(dat$armed)
```

body_camera is a binary variable that is True when news reports document that the officer was wearing a body camera and may have recorded at least some part of the interaction. The flee variable describes the manner in which a victim was moving away from the officers, if at all, and the responses are either 'Foot', 'Car', 'Not fleeing', 'Other', or blank. The armed variable describes if the victim possessed an object that the officer believed could harm others with responses of 'undetermined' (it is not know whether or not the victim was armed), 'unknown' (the victim possessed an object but what it was is not known), 'unarmed', or the identifier of the object itself.

d. What are three weapons that you are surprised to find in the "armed" variable? Make a table of the values in "armed" to see the options.

```
table(dat$armed)
```

```
##
##                                    air conditioner
##                         207                       1
##                  air pistol             Airsoft pistol
##                           1                       3
##                          ax                 barstool
##                          24                       1
##                baseball bat      baseball bat and bottle
##                          20                       1
## baseball bat and fireplace poker      baseball bat and knife
##                           1                       1
##                       baton                  BB gun
##                           6                      15
##           BB gun and vehicle             bean-bag gun
##                           1                       1
##                beer bottle               binoculars
##                           3                       1
##                blunt object                  bottle
##                           5                       1
##               bow and arrow               box cutter
##                           1                      13
##                       brick          car, knife and mace
##                           2                       1
##                     carjack                   chain
##                           1                       3
##                   chain saw                 chainsaw
##                           2                       1
##                       chair          claimed to be armed
##                           4                       1
##           contractor's level           cordless drill
##                           1                       1
##                    crossbow                  crowbar
##                           9                       5
##                   fireworks                 flagpole
##                           1                       1
##                  flashlight              garden tool
##                           2                       2
##                glass shard                   grenade
##                           4                       1
##                         gun              gun and car
##                        3798                      12
##               gun and knife          gun and machete
##                          22                       3
##               gun and sword          gun and vehicle
##                           1                      17
##          guns and explosives                  hammer
##                           3                      18
##                hand torch                 hatchet
##                           1                      14
##             hatchet and gun                 ice pick
##                           2                       1
##            incendiary device                    knife
```

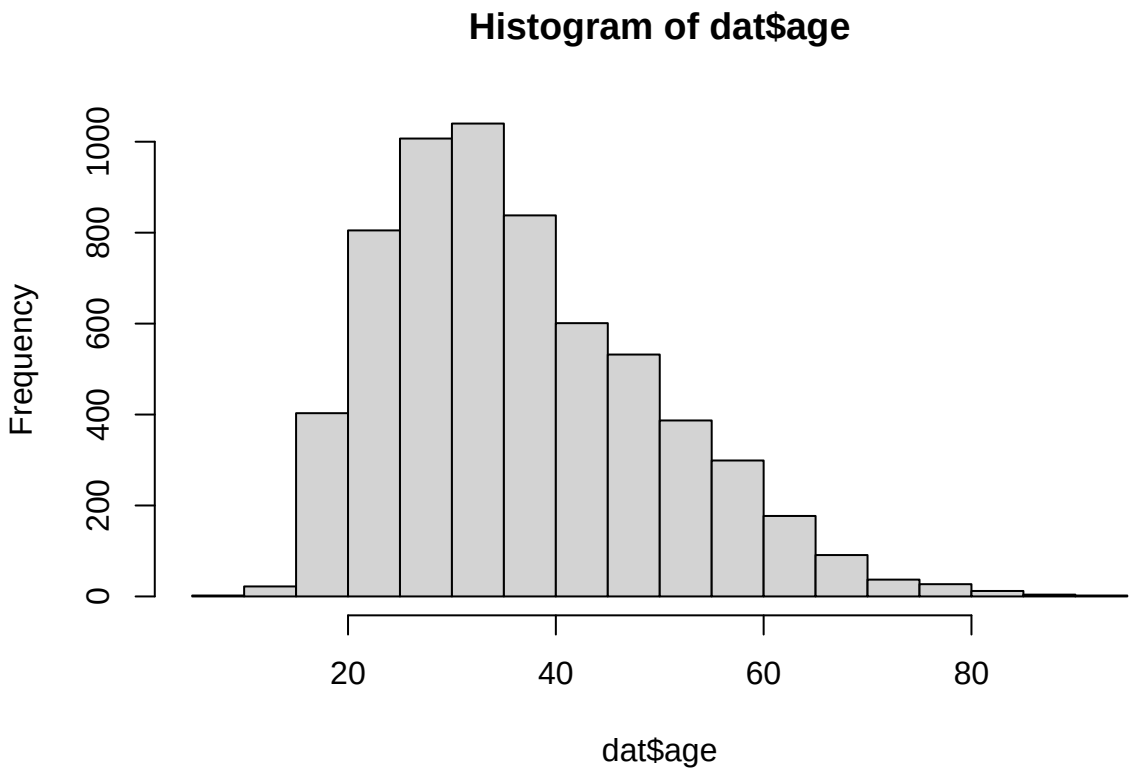
```
##                            2                          955
##                knife and vehicle              lawn mower blade
##                            1                            2
##                      machete                machete and gun
##                           51                            1
##                   meat cleaver              metal hand tool
##                            6                            2
##                   metal object                   metal pipe
##                            5                           16
##                     metal pole                   metal rake
##                            4                            1
##                    metal stick                   microphone
##                            3                            1
##                    motorcycle                     nail gun
##                            1                            1
##                            oar                   pellet gun
##                            1                            3
##                            pen                  pepper spray
##                            1                            2
##                      pick-axe                piece of wood
##                            4                            7
##                           pipe                    pitchfork
##                            7                            2
##                           pole               pole and knife
##                            3                            2
##                 railroad spikes                         rock
##                            1                            7
##                  samurai sword                    scissors
##                            4                            9
##                    screwdriver                  sharp object
##                           16                           14
##                         shovel                        spear
##                            7                            2
##                         stapler          straight edge razor
##                            1                            5
##                          sword                       Taser
##                           23                           34
##                       tire iron                 toy weapon
##                            4                          226
##                        unarmed                undetermined
##                          421                          188
##                 unknown weapon                      vehicle
##                           82                          213
##               vehicle and gun          vehicle and machete
##                            8                            1
##                 walking stick                   wasp spray
##                            1                            1
##                        wrench
##                            1
```

Flashlight, wasp spray, and ice pick.

Problem 2 (10 points)

a. Describe the age distribution of the sample. Is this what you would expect to see?

```
hist(dat$age)
```




```
table(dat$age)
```

```
##
##   6  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30
##   2   1   2   3  16  35  56 109  98 105 124 138 148 179 216 188 217 194 204 204
##  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50
## 223 207 205 209 196 186 183 164 165 140 142 112 120 101 126 109 115 110 103  95
##  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70
##  89  82  78  71  67  72  58  56  66  47  44  45  35  25  28  19  23  16  17  16
##  71  72  73  74  75  76  77  78  79  80  81  82  83  84  86  88  89  91
##  12   7   7   6   5  12   5   1   4   5   3   2   3   4   2   1   1   2
```

The age distribution is skewed slightly to the right, with the largest group of victims being in their mid-late 20s to early-mid 30s. There are very few victims below the age of 15 and above the age of 71, however the tail on the right is longer. This is what I expected to see as the largest frequency shown here are part of the population that has the largest interaction with the police, however I was a little surprised by how few older teenage fatal shootings there were based on the increased involvement in crime and dangerous behavior that is demonstrated by this age group. I was also surprised by how high some of the values were on the right side of the peak.

b. To understand the center of the age distribution, would you use a mean or a median, and why? Find the one you picked.

```
summary(dat$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    6.00   27.00   35.00   37.12   45.00   91.00     308
```

```
median(dat$age, na.rm = TRUE)
```

```
## [1] 35
```

As the age distribution is not symmetrical, the mean and median do not coincide. In this case, the distribution is skewed so the mean will be further towards the the tail, while the median will better represent the central tendency of the data. The median of age is 35 (excluding observations where age was not included).

c. Describe the gender distribution of the sample. Do you find this surprising?

```
table(dat$gender)
```

```
##
##         F    M
##    3  293 6298
```

There were 6298 male victims and 293 female victims, making almost 95.5% of all victims of police shooting being male. While I expected there to be significantly more male than female victims based on the gender division for criminal activity especially for violent crimes, however I did not expect the difference to be this large.

Problem 3 (10 points)

a. How many police officers had a body camera, according to news reports? What proportion is this of all the incidents in the data? Are you surprised that it is so high or low?

```
table(dat$body_camera)
```

```
##
## False  True
##  5684   910
```

According to news reports, in only 910 of the incidents, the police officers had a body camera. Therefore, in only 16% of all fatal shootings by police, the police had a body camera. I am surprised that it is so low, as the vast majority of officers that I see on a day-to-day basis are wearing them, but I have to imagine that much of my perception about body-worn cameras is based on where I have lived and the policies in these cities.

b. In how many of the incidents was the victim fleeing? What proportion is this of the total number of incidents in the data? Is this what you would expect?

```
table(dat$flee)
```

```
##
##                  Car         Foot Not fleeing        Other
##         491        1058         845        3952         248
```

In 1058 cases the victim was fleeing by car, in 845 cases the victim was fleeing on foot, in 3952 cases the victim was not fleeing, in 248 cases the value was listed as Other, and no response was recorded for 491 cases. Excluding the 491 cases and 248 responses for 'Other', out of the remaining 5855 cases only 1903 cases the victim flee-ed in. Therefore, in 32.5% of cases the victim flee-ed (not including observations where the response was blank or Other due to lack of information). This statistic is a little suprising to me as I expected the percentage of victims that were fleeing to be higher as it does have some relation to the threat they pose to the public, especially if they are armed.

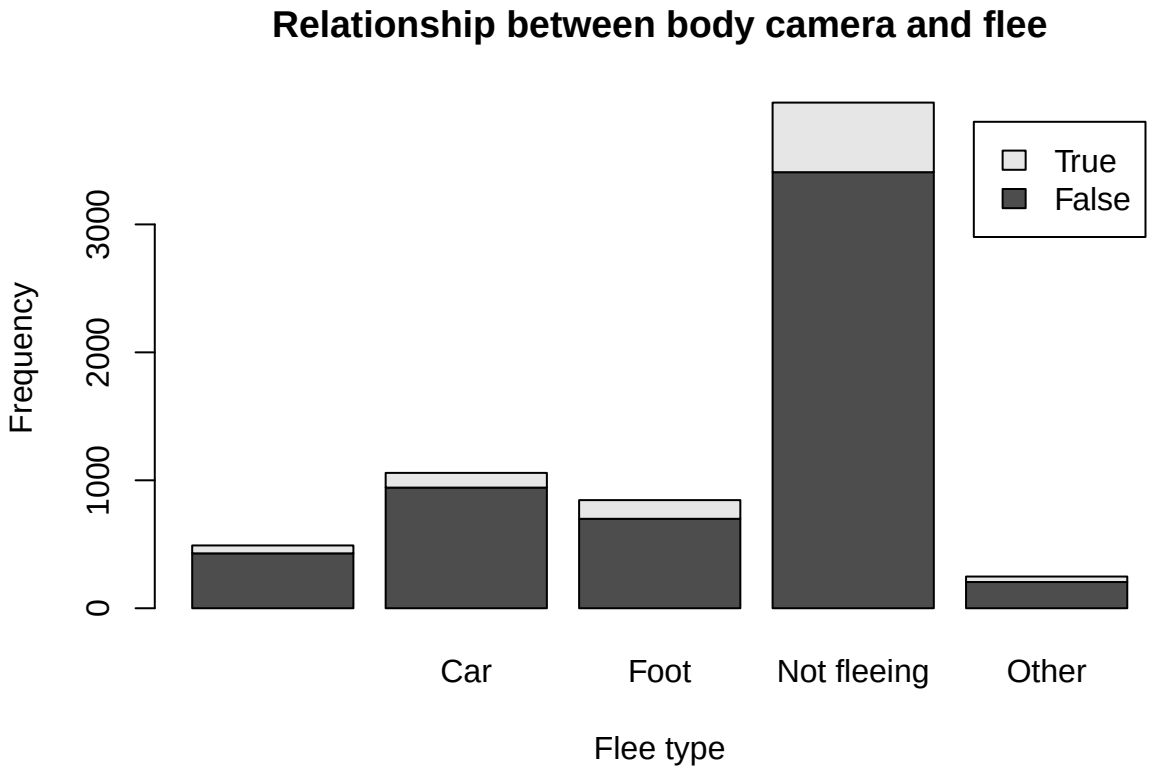
Problem 4 (10 points)

a. Describe the relationship between the variables "body camera" and "flee" using a stacked barplot. What can you conclude from this relationship?

Hint 1: The categories along the x-axis are the options for "flee", each bar contains information about whether the police officer had a body camera (vertically), and the height along the y-axis shows the frequency of that category).

Hint 2: Also, if you are unsure about the syntax for barplot, run ?barplot in R and see some examples at the bottom of the documentation. This is usually a good way to look up the syntax of R code. You can also Google it.

```
barplot(table(dat$body_camera, dat$flee), main = "Relationship between body camera and flee", xlab = "F
```



From this stacked barplot, there does not appear to be a significant relationship between the officer wearing a body-worn camera and whether/ how a victim flees. It shows that in all cases of fleeing and not fleeing, officers are much more likely to not be wearing a body-worn camera.While it appears that those not fleeing and the officer is wearing a body camera vs those fleeing and the officer wearing a body camera is larger, in actuality these ratios are very similar and just looks larger because of the size of the column.

Extra credit (10 points)

a. What does this code tell us?

```
mydates <- as.Date(dat$date)
head(mydates)
#tail(mydates)
(mydates[length(mydates)] - mydates[1])
```

This code first modifies the character representation of the date provided in the table and classifies it as an object of the Date class. The head function is used to show the first couple values in this table of format-modified dates. Lastly, the difference between the final date (same as table[length(table)]) included in the table and the first date in the table is found to determine the number of days between the two. This shows that this data represents 2458 days fro January 2nd 2015 to September 25 2021.

b. On Friday, a new report was published that was described as follows by The Guardian: "More than half of US police killings are mislabelled or not reported, study finds." Without reading this article now (due to limited time), why do you think police killings might be mislabelled or underreported?

As police killings reflect poorly on police officers and as it is their responsibility to report them, there is

a motivation and ability to not accurately report this data so that it presents a more positive view of the police/ their abilities. Additionally, it is possible that these incidents are being intentionally mislabled to better a precincts statistics or simply due to error caused by excessive movement of paperwork and lack of a streamlined process.

c. Regarding missing values in problem 4, do you see any? If so, do you think that's all that's missing from the data?

For the flee variable, there are 491 responses that left this column blank. I would expect there to be more missing values in the dataset. Other missing values include 3 victims that do not have a gender listed and 752 victims whose race value was missing.